Solomonoff Prediction and Occam's Razor

Sterkenburg, Tom

Published in:
Philosophy of Science

DOI:
10.1086/687257

Solomonoff Prediction and Occam's Razor

Tom F. Sterkenburg*†

Algorithmic information theory gives an idealized notion of compressibility that is often presented as an objective measure of simplicity. It is suggested at times that Solomonoff prediction, or algorithmic information theory in a predictive setting, can deliver an argument to justify Occam's razor. This article explicates the relevant argument and, by converting it into a Bayesian framework, reveals why it has no such justificatory force. The supposed simplicity concept is better perceived as a specific inductive assumption, the assumption of effectiveness. It is this assumption that is the characterizing element of Solomonoff prediction and wherein its philosophical interest lies.

1. Introduction. *Occam's razor* is the principle in science that tells us to prefer the simplest available hypothesis that fits the data. As a pragmatic principle, it might strike one as obvious, but it is often interpreted in a stronger fashion. As an epistemic principle, Occam's razor comes with a promise that a preference for simpler hypotheses is somehow more likely to lead us to the truth. This raises the difficult question of how to ground such a promise, thus, to justify the epistemic principle. Still before this is the nontrivial problem of how to actually measure simplicity.

Algorithmic information theory, also known as *Kolmogorov complexity* after Kolmogorov (1965), is sometimes believed to offer us a general and objective measure of simplicity. The idea is that a data object, like the specification of a hypothesis, is simpler as it is more compressible, meaning that we can capture it in a shorter description. With the aid of the theory of computability, this idea can be made formally precise, culminating in the defi-

Received May 2015; revised September 2015.

*To contact the author please write to: Algorithms and Complexity Group, Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands; e-mail: tom@cwi.nl.

†For valuable feedback on several versions and presentations of this article, I am indebted to Peter Grünwald, Jan-Willem Romeijn, the members of the Groningen PCCP seminar, Simon Huttegger, Hannes Leitgeb, Samuel Fletcher, Filippo Massari, Teddy Seidenfeld, and an anonymous referee. This research was supported by NWO Vici project 639.073.904.

Philosophy of Science, 83 (October 2016) pp. 459–479. 0031-8248/2016/8304-0001$10.00

nition of a data object's Kolmogorov complexity as the length of its shortest description. In the standard textbook on the subject, we read: "This gives an objective and absolute definition of 'simplicity' as 'low Kolmogorov complexity.' Consequently, one obtains an objective and absolute version of the classic maxim of William of Ockham" (Li and Vitányi 2008, 260).

But this is not all. The first variant of Kolmogorov complexity to appear in the literature, by the hand of Solomonoff (1960, 1964), was part of a theory of prediction. Solomonoff's central achievement was the definition of an idealized method of prediction that employs this complexity measure to give greater probability to simpler extrapolations of past data. Moreover, Solomonoff (1978) was able to formally prove that this prediction method is reliable in the sense that it will generally lead us to the truth.

Here emerges an argument that is suggested in many writings on the subject. The argument concludes from (1) the definition of a type of predictor with a preference for simplicity and (2) a proof that predictors of this type are reliable that (per Occam's razor) a preference for simplicity will generally lead us to the truth. Thus, remarkably, it is an argument to justify Occam's razor.

In this article, I consider this argument in detail. The conclusion will be that it does not succeed. I reach this conclusion by employing a specific representation theorem to translate the argument in terms of Bayesian prediction. This translation reveals that the apparent simplicity bias is better understood as a particular inductive assumption, which by a basic property of Bayesian prediction methods entails reliability under that very same assumption—leaving the conclusion of the argument without justificatory force.

The main positive contribution of this article is the observation that—rather than simplicity—it is the assumption or constraint of effectiveness that is the central element of Solomonoff's theory of prediction. This can serve as the starting point for a more careful philosophical appraisal of Solomonoff's theory. While numerous substantial claims about the theory's philosophical merits have been advanced from the angle of theoretical computer science, attention in the philosophical literature has so far been largely restricted to the occasional mention in overview works. This is unfortunate. Not only can the theory be seen as the progenitor to multiple successful modern approaches in statistics and machine learning, including *universal prediction* or *prediction with expert advice* (see Cesa-Bianchi and Lugosi 2006) and the *principle of minimum description length* (MDL; see Rissanen 1989; Grünwald 2007); the theory itself originated as a branch of a major philosophical project—namely, Carnap's early program of inductive logic, pursued with tools from information theory and computability theory. In this capacity the theory brings together a diverse range of motifs from the philosophy of induction, which in turn connect the theory to several other approaches: among

those we find *formal learning theory* (see Kelly 1996), which likewise puts effectiveness center stage, and the project of *meta-induction* (Schurz 2008), the philosophical counterpart to prediction with expert advice. The broader aim of the current article is to convey this to the reader.

The plan is as follows. I start in section 2 with covering some essential preliminaries on sequential prediction and effectiveness. In section 3, I present the details of the argument to justify Occam's razor. Section 4 introduces Bayesian prediction. Section 5, which forms the philosophical heart of the article, is devoted to a representation theorem that bridges Solomonoff's predictors and Bayesian prediction. In section 6, I employ this representation theorem to translate the argument in terms of Bayesian prediction, thereby revealing the hidden assumptions and showing why the argument fails to deliver a justification of Occam's razor. I conclude in section 7.[1]

2. Setting the Stage. Here, I introduce the minimal amount of terminology and notation that we need in this article. Section 2.1 covers sequential prediction; section 2.2 covers computability and effectiveness.

2.1. Sequential Prediction. We consider sequential prediction of binary digits (*bits*), elements of the set $\mathbb{B} := \{0, 1\}$. Having witnessed a finite sequence σ of bits, we are to make a probability forecast, based on σ only, of what bit comes next; then this bit is revealed, and the procedure is repeated. Dawid (1984) names it the *prequential* approach, for sequential prediction in a probabilistic fashion.

Sources and Predictors. A *probabilistic source* represents a random bit generating process. It is a function that returns for every finite se-

1. This article concerns the justification of Occam's razor in the approach to predictive inference on the basis of algorithmic information theory, the approach invented by Solomonoff. It is important to note that I make no claims here about other approaches to statistical inference in the field (like the *Kolmogorov structure function*; see Vitányi 2005) and that my observations certainly have no direct bearing on approaches that are for large part only inspired by algorithmic information theory, like the MDL principle. While Rissanen (1989) acknowledges that his "main source of inspiration in developing the *MDL* principle for general statistical problems has been the theory of *algorithmic* complexity [algorithmic information theory]," he is quick to add that "the role of the algorithmic complexity theory is inspirational, only, for almost everything about it, such as the idea of a model and even the very notion of complexity, must be altered to make the ideas practicable" (10). Two relevant ways in which the theory is fundamentally different are that the notion of complexity in the MDL approach pertains to hypothesis classes or models (so that the resulting simplicity bias is akin to that in the *Bayes factor* method; see Kass and Raftery 1995), rather than to data sequences or single hypotheses as in Solomonoff's theory (Grünwald 2007, 31), and that effectiveness plays no fundamental part, whereas this is the key ingredient in Solomonoff's theory.

quence of outcomes the probability that this sequence is generated. Hence, it is a function $S : \mathbb{B}^* \to [0, 1]$ on the set $\mathbb{B}^* \coloneqq \cup_{t\in\mathbb{N}} \mathbb{B}^t$ of all finite outcome sequences such that, first, the initial probability equals 1 (so $S(\epsilon) = 1$ for the empty sequence ϵ), and, second, it fulfills a condition of compatibility: for all sequences σ, the summed probability of both 1-bit extensions of σ equals the probability of σ (so $S(\sigma 0) + S(\sigma 1) = S(\sigma)$).

A basic example of a probabilistic source is the *Bernoulli* source with parameter p, which corresponds to the process of repeatedly generating a bit with the same probability p of outcome 1. The special case of the Bernoulli source corresponding to the process of repeatedly generating a bit with both outcomes having equal probability is given by $S(\sigma) \coloneqq 2^{-|\sigma|}$, where $|\sigma|$ denotes the length of sequence σ. Another example is a *deterministic* source that just generates (say) the real number π in binary: it is defined by $S(\sigma) = 1$ for all those σ that are initial segments of the binary development of π, and $S(\sigma) = 0$ for all other sequences.

A *prediction method*, or simply *predictor*, is a function that returns for every finite sequence of outcomes a specific prediction. A prediction can be a single element 0 or 1, but we take it more generally as a probability distribution over both possibilities. Thus, a predictor is a function $P : \mathbb{B}^* \to \mathcal{P}_{\mathbb{B}}$, with $\mathcal{P}_{\mathbb{B}}$ the class of all probability distributions over 0 and 1.

As an example, analogous to the Bernoulli source, one can define a predictor that always returns the probability distribution assigning probability p to outcome 1. Another example is the "maximum likelihood" predictor that returns for sequence σ the probability distribution that assigns to outcome 1 the relative frequency of 1's in σ.

As suggested by the Bernoulli example, a probabilistic source determines in a straightforward way a predictor, and vice versa. So, importantly, we can treat predictors and probabilistic sources as formally interchangeable objects. In all of what follows, I use the term "probabilistic source" in an interpretation-neutral way, to simply refer to a function with the above formal properties. That way, it makes sense to define a probabilistic source and then interpret it as a predictor. Whenever I intend the interpretation of a probabilistic source as giving the objective probabilities or chances in a random process, I make this more explicit by talking about a data-generating probabilistic source.[2]

2. For clarity of presentation, I have taken some amount of liberty in simplifying notions and notation. Perhaps the most significant technical aspect that I ignore in the main text is that the notion of a probabilistic source is actually understood in a somewhat weaker sense. Namely, only the inequalities $S(\epsilon) \leq 1$ and $S(\sigma 0) + S(\sigma 1) \leq S(\sigma)$ for all σ are required of a source S: such a source is called a *semimeasure* in the algorithmic information theory literature (as opposed to a measure that satisfies the equalities). Also see n. 10.

Risk of a Predictor. To evaluate the anticipated performance of a predictor, we need a notion of *expected prediction error*. We take the expectation according to some presupposed actual data-generating probabilistic source S^*.

For the specification of prediction error, we have the choice of various functions of the error (or loss) of a prediction with respect to an actual outcome. Let us fix as our loss function the customary function in the area, the *logarithmic loss*; from this loss function we then obtain a specific measure $\mathrm{R}_t(S^*, P)$ of expected prediction error, or simply *risk*, of predictor P for the tth prediction.[3] By summing the risks over all instances $t \in \mathbb{N}$, we have a specification of the total risk $\mathrm{R}(S^*, P) := \Sigma_t \mathrm{R}_t(S^*, P)$.

2.2. Computability and Effectiveness. This subsection introduces the basic notions from the theory of computability that we require in our setting of sequential prediction.[4]

Turing Machines and Computability. A *Turing machine* represents a particular algorithm, or computer program. We need not be concerned here with the formal definition: think of a Turing machine M as a black box that, when presented with a bit sequence ρ for input, starts calculating and either halts at some point (producing a bit sequence σ for output: we write $M(\rho) = \sigma$) or goes on calculating forever.

The generally accepted *Church-Turing Thesis* states that every possible algorithm corresponds to some Turing machine. If a Turing machine represents a particular computer program, a universal Turing machine represents a general-purpose computer. A machine of this kind is called universal because it can emulate every other Turing machine. The reason that we can define such a machine is that it is possible to enumerate a list $\{M_i\}_{i \in \mathbb{N}}$ of all Turing machines in a calculable way, meaning that there is an algorithm that given an index j reconstructs the jth Turing machine from this list. A universal machine U implements such an algorithm: given as input the concatenation of a code sequence $\langle j \rangle$ for M_j and a sequence ρ, it will recon-

3. The logarithmic loss of a prediction giving probability p to actual outcome b is $-\log_2 p$. For the agreeable properties of the logarithmic loss function in "pure inference," see Bernardo and Smith (1994, 69–81) and Merhav and Feder (1998, 2127–28). The resulting risk function is the *Kullback-Leibler divergence*, or *relative entropy*, that has an interpretation as the quantity of information lost when estimating with the one source rather than the other. Nothing in our discussion hinges on this interpretation, and indeed not much hinges on the particular risk function: the relevant theorems 1 and 2 below (at least for sources that are measures) continue to hold for other standard risk functions, like the mean squared error. See Solomonoff (1978, 426–27) and Li and Vitányi (2008, 352–55).

4. For a much fuller treatment of the theory of computability in the context of algorithmic information theory, see Nies (2009) and Downey and Hirschfeldt (2010).

struct M_j and calculate $M_j(\rho)$. In symbols, $U(\langle i \rangle \rho) = M_i(\rho)$ for all $i \in \mathbb{N}$, $\rho \in \mathbb{B}^*$.

The Church-Turing Thesis is therefore equivalent to the statement that everything that is calculable is calculable by a universal Turing machine. I reserve the term *computable* for the formal property of being calculable by a universal Turing machine. Then a more economical formulation of the Church-Turing Thesis reads: everything that is calculable is computable.

Effectiveness. Let D stand for an arbitrary but countable set of elements. If D is not $\mathbb{B}^*$, I still simply write $M(d)$ and speak of machine M with input $d \in D$, where it is actually more proper to write $M(\langle d \rangle)$ and speak of M with input some code sequence $\langle d \rangle \in \mathbb{B}^*$ for d. This is possible because the elements of a countable set D can always be encoded by finite bit sequences.

Call a function $f : D \to D'$ *computable* if there exists a Turing machine M that represents it: $M(d) = f(d)$ for all $d \in D$. This definition applies to integer- and rational-valued functions. A real-valued function $f : D \to \mathbb{R}$ we call *computable* if some Turing machine can approximate its values up to an arbitrary level of precision: there is some computable rational-valued function $g : D \times \mathbb{N} \to \mathbb{Q}$ such that the difference $|f(d) - g(d, k)| < 1/k$ for all $k \in \mathbb{N}$. A somewhat weaker requirement than full computability is semicomputability. Call a function f *semicomputable* (from below) if some universal machine can compute ever-closer lower approximations to its values (without revealing how close). That is, for such f there exists a computable $g : D \times \mathbb{N} \to \mathbb{Q}$ with the property that for all $d \in D$ and all $s \in \mathbb{N}$ we have that $g(d, s) \leq g(d, s + 1)$ and $\lim_{s \to \infty} g(d, s) = f(d)$.

I will treat semicomputability as the minimal level of calculability and from this point on use the term *effective* for any function that satisfies it. Note that, since they are functions on bit sequences, we can directly apply this requirement to probabilistic sources and, hence, predictors.

Indeed, one might consider it a most basic requirement on what would still count as a predictor that it provides probability assessments that are at least in principle approximable by our means of calculation. With the Church-Turing Thesis, this means that the class of possible predictors should be restricted to the effective ones as defined above. This is a philosophical point that I return to in section 7.

Conversely, we accept that any effective predictor does represent a possible method of prediction. We must do so, if we are to grant that Solomonoff's predictor below indeed represents a method of prediction, or the argument is discredited from the start.

3. The Argument to Justify Occam's Razor. Li and Vitányi (2008) write, "It is widely believed that the better a theory compresses the data concerning some phenomenon under investigation, the better we have learned and generalized, and the better the theory predicts unknown data, following

the Occam's razor paradigm about simplicity. This belief is vindicated in practice but apparently has not been rigorously proved before. . . . We . . . show that compression is almost always the best strategy . . . in prediction methods in the style of R.J. Solomonoff" (347–48). The general form of the argument that we can distill from these words is as follows. First, we identify a class $\mathcal{Q}$ of predictors that have a distinctive preference for simplicity (predictors "following the Occam's razor paradigm"). Here I mean "distinctive," to convey that predictors outside $\mathcal{Q}$ do not possess such a simplicity bias. Second, we prove that these predictors are reliable ("almost always the best strategy").[5] Taken together, the two steps establish a connection between two seemingly distinct properties of a predictor: a preference for simplicity, on the one hand, and a general reliability, on the other. More precisely, the two steps together yield the statement that if a predictor possesses a simplicity bias, then it is reliable. Equivalently, predictors that possess a simplicity bias are reliable.

In short, the argument is as follows:

1. Predictors in class $\mathcal{Q}$ possess a distinctive simplicity bias.
2. Predictors in class $\mathcal{Q}$ are reliable.
∴ Predictors that possess a simplicity bias are reliable.

Occam's razor, in our setting of sequential prediction, is the principle that a predictor should possess a simplicity bias. The conclusion of the above argument provides an epistemic justification for the principle of Occam's razor, so stated. A predictor should possess a simplicity bias because if it does, it is reliable.[6]

We now need to make precise the two steps of the argument, including the relevant notions of simplicity and reliability. I discuss the explication of step 1 in section 3.1 and that of step 2 in section 3.2. I revisit the complete argument in section 3.3.

3.1. Step 1: The Predictor. A *monotone* machine is a particular kind of Turing machine that can be seen to execute an "online" operation: in the course of processing a continuous stream of input bits, it produces a potentially infinite stream of output bits. Formally, such a machine M has the property that for any extension ρ' of any input sequence ρ (we write $\rho \preccurlyeq \rho'$), if M

5. This phrasing suggests a weaker property than reliability (i.e., convergence to the truth), namely, *optimality* (convergence to predictions that are at least as good as those of any other prediction method). However, the proof that is referred to is about reliability.

6. Note that the proposed justification only asserts that a simplicity bias is sufficient for reliability. One might feel that a true justification should also include the necessity of a simplicity bias for reliability: only the predictors in $\mathcal{Q}$ are reliable. It is possible to revise the argument to yield the stronger statement. However, for ease of presentation I here stick to the former argument.

yields an output on ρ' at all, it yields an output $M(\rho')$ that is also an extension of $M(\rho)$ (so $M(\rho) \preccurlyeq M(\rho')$). The monotone machine model suffers no loss of generality in what can be computed: every function calculable by a standard Turing machine is calculable by some monotone machine. The reason why this machine model is central to the theory is that the monotonicity property allows us to directly infer from each monotone machine a particular probabilistic source. We now proceed to do so for a *universal* monotone machine.

So suppose we have some universal monotone machine U, and suppose we feed it random bits for input: we repeatedly present it a 0 or a 1 with equal probability 0.5. For any sequence ρ, the probability that we in this way end up giving the machine a sequence starting with ρ only depends on the length $|\rho|$: this probability is $2^{-|\rho|}$. Having processed ρ, the machine will have produced some output sequence. For a sequence σ of any length that starts this output sequence, we can say that input ρ has served as an instruction for U to produce σ. For this reason we call sequence ρ a *U-description* of σ.

We can now ask the question: If we feed machine U random bits, what is the probability that it will return the sequence σ? In other words, if we generate random bits, what is the probability that we arrive at some U-description of given σ? This probability is given by Solomonoff's *algorithmic* probabilistic source.

> **Definition 1** (Solomonoff 1964). The *algorithmic probabilistic source* $Q_U : \mathbb{B}^* \to [0, 1]$ via universal monotone Turing machine U is given by $Q_U(\sigma) \coloneqq \Sigma_{\rho \in D_{U,\sigma}} 2^{-|\rho|}$, with $D_{U,\sigma}$ the set of minimal U-descriptions of σ, that is, the set of sequences ρ such that $U(\rho) \succcurlyeq \sigma$ and not $U(\rho') \succcurlyeq \sigma$ for any shorter sequence $\rho' \prec \rho$.

We see that a sequence σ receives greater algorithmic probability $Q_U(\sigma)$ as it has shorter descriptions ρ. The accompanying intuition is that σ receives greater algorithmic probability as it is more compressible. If we further accept this measure of compressibility as a general measure of simplicity of finite data sequences, then we can say that a sequence receives greater algorithmic probability as it is simpler.

By the formal equivalence of probabilistic sources and predictors (sec. 2.1), we can reinterpret an algorithmic probabilistic source as an *algorithmic probability predictor*. Given data sequence σ, the probability according to predictor Q_U of bit b showing next is the conditional probability $Q_U(b \mid \sigma) \coloneqq Q_U(\sigma b)/Q_U(\sigma)$.

Following the above intuition about data compression, the one-bit extension σb with the greatest algorithmic probability $Q_U(\sigma b)$ among the two possibilities $\sigma 0$ and $\sigma 1$ is the one that is the more compressible. Consequently, we see from the above equation that $Q_U(b \mid \sigma)$ is greatest for the b such that

σb is the more compressible. Hence, the algorithmic probability predictor Q_U will prefer the bit b that renders the complete sequence σb more compressible. This is, in the words of Ortner and Leitgeb (2011, 734), "evidently an implementation of Occam's razor that identifies simplicity with compressibility."

The above reasoning applies to the algorithmic probability predictor Q_U for any choice of universal Turing machine U. Since there are infinitely many universal machines, we have an infinite class of algorithmic probability predictors.

Let us denote $\mathcal{Q} \coloneqq \{Q_U\}_U$ the class of algorithmic probability predictors via all universal machines U. Thus, we have specified a class of predictors $\mathcal{Q}$ that possess a distinctive simplicity-qua-compressibility bias.

3.2. Step 2: The Reliability of the Predictor. The crucial result is that under a "mild constraint" on the presupposed actual data-generating source S^*, we can derive a precise constant upper bound on the total risk of the predictor Q_U. The "mild constraint" on S^* is that it is itself effective. It can be shown that this property guarantees that we can represent S^* in terms of the behavior of some monotone machine M^*, which in turn can be emulated by the universal monotone machine U. Then we can define a weight $W_U(S^*)$ that is a measure of how easily U can emulate M^*, more precisely, how short the U-codes of M^* are. This weight gives the constant bound on Q_U's risk, as defined in section 2.1.

> **Theorem 1** (Solomonoff 1978). For every effective data-generating probabilistic source S^*, and for every universal monotone machine U, $\mathrm{R}(S^*, Q_U) \leq -\log_2 W_U(S^*)$.

A direct consequence of the constant bound on the total risk is that the predictions of Q_U must rapidly converge, with S^*-probability 1, to the presupposed actual probability values given by probabilistic source S^*.[7] With S^*-probability 1, we have that $Q_U(b|X^{t-1}) \overset{t\to\infty}{\to} S^*(b|X^{t-1})$.[8]

7. The type of convergence of theorem 1, called *convergence in mean sum* to a constant by Hutter (2003), lies between the type of convergence results that are silent about the rate of convergence (like the merger-of-opinion results of Blackwell and Dubins [1962] and Gaifman and Snir [1982]) and the type of results that provide an explicit bound on the risk for the tth prediction. Convergence in mean sum to a constant is a fairly strong kind of convergence: common bounds on the risk for the tth prediction in results of the second type cannot guarantee a constant bound on the total risk. This warrants speaking of "rapid" convergence. However, the bound of theorem 1 becomes less surprising if one realizes that the class of possible effective data-generating sources is only countable (as a result of effectiveness), whereas convergence results normally presuppose uncountable hypothesis classes (cf. Solomonoff 1978, 427; Li and Vitányi 2008, 357–58).

8. Strictly speaking, this probability 1 convergence can only hold for (and theorem 1 is in the literature only stated for) sources S^* that are measures (see n. 2). However, the-

Let us more precisely define a predictor P to be reliable for S^* if, with S^*-probability 1, its predictions converge to the actual conditional S^*-probability values. Then, under the "mild constraint" of effectiveness of the actual source, the predictor Q_U is reliable. It would motivate the conclusion that Q_U is reliable "in essentially every case."

3.3. The Complete Argument. Let me restate the full argument:

1. Predictors in class $\mathcal{Q}$ possess a distinctive simplicity-qua-compressibility bias.
2. Predictors in class $\mathcal{Q}$ are reliable in essentially every case.

$\therefore$ Predictors that possess a simplicity-qua-compressibility bias are reliable in essentially every case.

Again, the conclusion of the argument asserts a connection between two seemingly distinct properties of predictors: a preference for simplicity and a general reliability. The establishment of this connection between a simplicity preference and a general reliability justifies the principle that a predictor should prefer simplicity, the principle of Occam's razor.

Note, however, that compared to the statement of the argument at the beginning of this section, I have added a minor qualification to both of the steps. Both qualifications are actually very much related, and spelling them out will show that the two properties are not so distinct after all. At heart, it is this fact that makes the conclusion of the argument fail to justify Occam's razor. In order to make all of this explicit, I now turn to the framework of Bayesian prediction.

4. Bayesian Prediction. Here, I discuss the definition and interpretation of Bayesian predictors (sec. 4.1), their reliability property of consistency (sec. 4.2), and the special class of effective Bayesian predictors (sec. 4.3), still in the setting of sequential bit prediction.

4.1. Bayesian Predictors. Bayesian prediction sets off with the selection of a particular class $\mathcal{S}$ of probabilistic sources, which serves as our class of hypotheses. (I here restrict discussion to hypothesis classes that are countable.) Next, we define a *prior distribution* (or *weight function*) $W : \mathcal{S} \to [0, 1]$ over our hypothesis class. The prior W is to assign a positive weight to the hypotheses and only the hypotheses in $\mathcal{S}$. An equally valid way of looking at things is that the definition of a particular prior W induces a hy-

orem 1 and a suitably analogous convergence are straightforwardly obtained for the general case of semimeasures.

pothesis class $\mathcal{S}$, simply defined as the class of hypotheses that receive a positive prior. In any case, we have $W(S) > 0 \Leftrightarrow S \in \mathcal{S}$.

Following Howson (2000) and Romeijn (2004), the prior embodies our *inductive assumption*. If induction, in our setting of sequential prediction, is the procedure of extrapolating a pattern in the past to the future, then the lesson of the new riddle of induction (Goodman 1955; also see Stalker 1994) is that there is actually always a multitude of candidate patterns. One can therefore only perform induction relative to a particular pattern or a hypothesis that represents a pattern that we have seen in the past data and that we deem projectable in the future. From a Bayesian perspective, the hypotheses that we give a positive prior represent the potential patterns in the data that we deem projectable; the other hypotheses, receiving prior 0, represent the patterns that we exclude from the outset. The great merit of the Bayesian framework is that it locates our inductive assumption very precisely, namely, in the prior.

We are now in the position to define a *Bayesian prediction method*. It is a prediction method that operates under the inductive assumption of the corresponding prior W.

> **Definition 2.** The *Bayesian predictor* $P_W^{\mathcal{S}} : \mathbb{B}^* \to [0, 1]$ via prior W on countable hypothesis class $\mathcal{S}$ is given by $P_W^{\mathcal{S}}(\sigma) := \Sigma_{S \in \mathcal{S}} W(S) S(\sigma)$.

Given data sequence σ, the probability according to $P_W^{\mathcal{S}}$ of bit b appearing next is the conditional probability $P_W^{\mathcal{S}}(b|\sigma) = P_W^{\mathcal{S}}(\sigma b)/P_W^{\mathcal{S}}(\sigma) = \Sigma_{S \in \mathcal{S}} W(S|\sigma) S(b|\sigma)$.

4.2. The Consistency of Bayesian Predictors. To operate under a particular inductive assumption means to predict well whenever the data stream under investigation follows a pattern that conforms to this inductive assumption. More precisely, if a Bayesian predictor operates under a particular inductive assumption, embodied by a prior over a particular hypothesis class $\mathcal{S}$, it will predict well whenever some hypothesis $S \in \mathcal{S}$ fits the data stream well: whenever the data stream is probable according to some $S \in \mathcal{S}$. More precisely still, the predictor will from some point on give a high probability to each next element of a data stream whenever there is some $S \in \mathcal{S}$ that has done and keeps on doing so.

This property is closely related to the property of predicting well whenever the data are in fact generated by some source $S^* \in \mathcal{S}$. If by "predicting well" we mean converging (with probability 1) to the true conditional probabilities, then this is again the property of reliability (sec. 3.2).

We can prove that any Bayesian predictor, operating under the inductive assumption of $\mathcal{S}$, is reliable under the assumption that the data are indeed generated by some source $S^* \in \mathcal{S}$. Indeed, we can derive a result completely

parallel to theorem 1 on the total risk (as defined in sec. 2.1) of the Bayesian predictors:[9]

> **Theorem 2.** For every data-generating probabilistic source $S^* \in \mathcal{S}$, and for every prior W on $\mathcal{S}$, $\mathrm{R}(S^*, P_W^{\mathcal{S}}) \leq -\log_2 W(S^*)$.

Again, this bound on the total risk of $P_W^{\mathcal{S}}$ entails its convergence to S^*. For every actual S^* that is indeed a member of the hypothesis class $\mathcal{S}$, the predictions of the Bayesian predictors $P_W^{\mathcal{S}}$ will converge with S^*-probability 1 to the actual probability values given by S^*. This is called the *consistency* property of Bayesian predictors.

4.3. The Effective Bayesian Predictors. Recall that the second step of the argument for Occam's razor relied on the "mild assumption" of effectiveness. This is an inductive assumption. In the Bayesian framework, we can explicitly define the class of predictors that operate under this inductive assumption.

Let $\mathcal{S}_{\mathrm{eff}}$ be the class of probabilistic sources that are effective. The inductive assumption of effectiveness is expressed by any prior W that assigns positive weight to the elements and only the elements of this class. If we moreover put the constraint of effectiveness on the prior W itself, the resulting Bayesian mixture predictor $P_W^{\mathcal{S}_{\mathrm{eff}}}$ will itself be effective. A predictor of this kind we call an *effective Bayesian mixture predictor*, or effective mixture predictor for short.[10]

> **Definition 3.** The *effective Bayesian mixture predictor* $P_W^{\mathrm{eff}} : \mathbb{B}^* \to [0, 1]$ via effective prior W on $\mathcal{S}_{\mathrm{eff}}$ is given by $P_W^{\mathrm{eff}}(\sigma) \coloneqq P_W^{\mathcal{S}_{\mathrm{eff}}}(\sigma) = \Sigma_{S \in \mathcal{S}_{\mathrm{eff}}} W(S) S(\sigma)$.

Let $\mathcal{R} \coloneqq \{P_W^{\mathrm{eff}}\}_W$ denote the class of effective mixture predictors via all effective priors W, that is, the class of all effective mixture predictors.

9. This "folklore" result (see, e.g., Barron 1998) could also be attributed to Solomonoff (1978), as it follows from the exact same proof as the one for theorem 1. See Poland and Hutter (2005). Here, too, the qualification of n. 8 applies.

10. The class of effective sources (semicomputable semimeasures; see n. 2) was first described by Zvonkin and Levin (1970), although Solomonoff (1964) already indicated a mixture over the class of computable measures. The shortcoming of the class of computable measures is that it cannot be computably enumerated; consequently, a mixture predictor $P_W^{\mathcal{S}_{\mathrm{comp}}}$ cannot be effective. In contrast, the larger class $\mathcal{S}_{\mathrm{eff}}$ of semicomputable semimeasures can be enumerated, and the mixture $P_W^{\mathcal{S}_{\mathrm{comp}}}$ is effective (as long as W is). (Since for measures, semicomputability already implies full computability, the weakening to semicomputability necessitates the weakening to semimeasures.) This can be seen as the motivation for introducing the class of probabilistic sources corresponding to the semicomputable semimeasures, rather than the seemingly more natural class of computable measures.

5. The Representation Theorem. Theorem 3 below is a *representation theorem* that forms the bridge between the algorithmic probability predictors and the effective mixture predictors, and that is the key to defusing the argument to justify Occam's razor in section 6. After the statement of the theorem in section 5.1, I discuss how the theorem illuminates a central theme surrounding Solomonoff's algorithmic probabilistic source, namely, its claim to objectivity. I initiate this discussion in section 5.2 on Solomonoff's original motivation to define an objective-logical measure function in the spirit of Carnap and complete it in section 5.3 on the correspondence between the choice of universal machine and the choice of Bayesian effective prior. Finally, section 5.4 treats the two directions of reading the theorem, in analogy to the original representation theorem of de Finetti.

5.1. The Theorem. The crucial fact is that definition 3 of the effective mixture predictor is equivalent to definition 1 of the algorithmic probability predictor. Recall that $\mathcal{Q} = \{Q_U\}_U$ denotes the class of algorithmic probability predictors via all universal monotone Turing machines U and that $\mathcal{R} = \{P_W^{\text{eff}}\}_W$ denotes the class of effective mixture predictors via all effective priors W. Then:

Theorem 3 (Wood, Sunehag, and Hutter 2013). $\mathcal{Q} = \mathcal{R}$.

Thus, every algorithmic probability predictor via some U is an effective mixture predictor via some W and vice versa.[11]

Among the philosophical fruits of theorem 3 is the light it sheds on the discussion about the element of subjectivity in the definition of the algorithmic probabilistic source. I spell this out in section 5.3; to prepare the ground I first discuss the origin of Solomonoff's work in Carnap's early program of inductive logic.

5.2. Algorithmic Probability as an Objective Prior. Solomonoff makes explicit reference to Carnap (1950) when he sets out his aim: "we want $c(a, T)$, the degree of confirmation of the hypothesis that [bit] a will follow, given the evidence that [bit sequence] T has just occurred. This corresponds to Carnap's probability$_1$" (1964, 2). Solomonoff's restriction of scope to what we have been calling sequential prediction aligns with Carnap's position that "predictive inference is the most important kind of inductive infer-

11. I should note that theorem 3 is established by a fairly simple derivation, and even the authors themselves consider it only a minor improvement on the well-known asymptotic equivalence of the members of $\mathcal{Q}$ and $\mathcal{R}$. The claim of this article is that the theorem presents (the sharpest expression of) a conceptually very significant fact about Solomonoff's theory.

ence" (1950, 568), from which other kinds may be construed as special cases. Carnap's "singular predictive inference," which is "the most important special case of the predictive inference" (568), concerns the degree of confirmation bestowed on the singular prediction that a new individual c has property M by the evidence that s_1 out of s individuals witnessed so far have this property M. If we translate this information into a bit sequence by simply writing 1 at the ith position for the ith individual having property M (and 0 otherwise), then we recover the problem of sequential prediction.[12]

Solomonoff's explication of degree of confirmation in our notation is the conditional algorithmic probability $Q_U(b|\sigma) = Q_U(\sigma b)/Q_U(\sigma)$, analogous to a Carnapian confirmation function $\mathfrak{c}$ that is defined by $\mathfrak{c}(h, e) \coloneqq \mathfrak{m}(h|e) = \mathfrak{m}(h \,\&\, e)/\mathfrak{m}(e)$ for an underlying regular measure function $\mathfrak{m}$ on sentences in a chosen monadic predicate language. To Carnap, the value $\mathfrak{m}(h)$ that equals the null confirmation $\mathfrak{c}_0(h)$ of h is "the degree of confirmation of h before any factual information is available" (1950, 308), which he allows might be called the "initial probability" or the "probability a priori" of the sentence. Degree of confirmation is a "logical, semantical concept" (19), meaning that the value $\mathfrak{c}(h, e)$ is established "merely by a logical analysis of h and e and their relations" (20), independent of any empirical fact, and so the underlying null confirmation $\mathfrak{c}_0$ also corresponds to "a purely logical function for the argument h" (308). Thus, $\mathfrak{c}_0$ is an objective prior distribution on sentences, where its objectivity derives from its logicality (43).

Likewise, Solomonoff seeks to assign "a priori probabilities" to sequences of symbols; although the approach he takes is to "examine the manner in which these strings might be produced by a universal Turing machine" (1964, 3), following an intuition about objectivity deriving from computation. The resulting explication of the null confirmation is the familiar algorithmic probabilistic source, which is indeed commonly referred to in the literature as the "universal a priori distribution" on the finite bit sequences.

5.3. The Element of Subjectivity. However, if the algorithmic probabilistic source is supposed to function as a "*single* probability distribution to use as the prior distribution in each different case" (Li and Vitányi 2008, 347), then it starts to look problematic that Q_U is not uniquely defined (cf. Solomonoff 1986, 477; Hutter 2007, 44–45).

12. Note that this translation presupposes a (temporal) ordering of individuals, which is something Carnap did not presuppose (1950, 62–64). This is an important deviation: for Solomonoff, sequences 00001111 and 10101010 are different and should presumably confer different degrees of confirmation on the next bit being 1; for Carnap both sentences, translated back, express the same fact that four individuals in a sample of eight have property M.

The Subjective Choice of Universal Machine. The fact is that the definition of the algorithmic probabilistic source retains an element of arbitrariness or subjectivity in the choice of universal machine U. There does exist an important *Invariance Theorem* to the effect that the shortest descriptions via one universal machine U are not more than a fixed constant longer than the shortest descriptions via another U'. This implies that the probability assignments of two algorithmic probability sources via different machines U and U' never differ more than a fixed factor, which in turn implies that any two different Q_U and $Q_{U'}$ converge to the same probability values as data sequences get longer: Machines Q_U and $Q_{U'}$ are asymptotically equivalent. The Invariance Theorem is generally taken to grant the definition of the algorithmic probability source a certain robustness. Indeed, the formulation of this theorem, independently by Solomonoff (1964), Kolmogorov (1965), and Chaitin (1969), is considered to mark the birth of algorithmic information theory. In Kolmogorov's own words, "The basis discovery . . . lies in the fact that the theory of algorithms enables us to limit this arbitrariness [of a complexity measure that depends on a particular description method] by the determination of a 'complexity' that is almost invariant" (quoted in Shiryaev 1989, 921; also see Li and Vitányi 2008, 95–99, 192).

However, the constant factor that binds two different sources can still be arbitrarily large. And there does not appear to be a principled way to single out a "most natural" or objective universal machine with which to define the algorithmic probabilistic source.[13]

The Shift to the Subjective. Carnap himself (1945, 1950), when he does propose as an explicatum of probability$_1$ a confirmation function $\mathfrak{c}^*$ based on a unique measure function $\mathfrak{m}^*$, is careful not to make the claim that "$\mathfrak{c}^*$ is a perfectly adequate explicatum of probability$_1$, let alone that it is the only adequate one" (1950, 563), and he indeed already (in Carnap 1952) resorts to a continuum of confirmation functions $\mathfrak{c}_\lambda$ parametrized by $\lambda \in [0, \infty]$. Undeniably, "the selection of a particular value of λ to uniquely determine a measure seems in the grand tradition of subjective theories of probability" (Suppes 2002, 198).[14]

The same can be said of the selection of a particular universal machine U to uniquely define Q_U, but acceptance of this circumstance has been slow in

13. Müller (2010) presents an interesting attempt to isolate a machine-invariant version of algorithmic probability. He concludes that "*there is no way to get completely rid of machine-dependence*, neither in the approach of this paper nor in any similar but different approach" (126).

14. Also see Jeffrey (1973, 302–3), for a brief and lucid evaluation of Carnap's subjectivism, and Zabell (2011, 301–5), for a more extensive overview of Carnap's "shift to the subjective."

the field: "for quite some time I felt that the dependence of [the algorithmic probabilistic source] on the reference machine was a serious flaw in the concept, and I tried to find some 'objective' universal device, free from the arbitrariness of choosing a particular universal machine" (Solomonoff 2009, 9–10). Nevertheless, in his later writings Solomonoff, too, turned away from the idea of a single most objective universal machine and came to embrace the choice of universal machine as an inevitable and essentially subjective element of prior information in the definition of the algorithmic probabilistic source (9–11).

The Subjective Choice of Effective Prior. The subjective element that lies in the choice of a specific universal machine is analogous to the subjective element in the choice of a specific effective prior in a Bayesian mixture over effective hypotheses. Note that the priors W and W' of any two Bayesian predictors P^{eff}_{W} and $P^{\text{eff}}_{W'}$ give positive weight to each other, which again implies that their probability assignments do not differ more than these weight factors, but, again, those weights may be arbitrarily small. Moreover, like universal machines, some effective priors appear more natural than others, and some complicated priors would probably look very unnatural, but there does not appear to be a principled way to single out a most natural or objective one.

A correspondence between the choice of universal machine and the choice of Bayesian prior over effective hypotheses has been noted before, for instance, by Wallace (2005, 401–4). Theorem 3 tells us that the analogy between universal monotone machines and effective priors over the effective probabilistic sources is in fact an exact correspondence.

5.4. Reading the Representation Theorem. De Finetti's celebrated representation theorem (1937/1964) states (translated to our setting of sequential bit prediction) the equivalence of a particular class of predictors, namely, those that are exchangeable (i.e., that assign the same probability to sequences with identical numbers of 0's and 1's), and a particular class of Bayesian mixtures, namely, those densities over the independently and identically distributed (i.i.d.) sources. Theorem 3 likewise states the equivalence of a particular class of predictors, the class of algorithmic probability predictors, and a particular class of Bayesian mixtures, the effective mixtures over the effective sources.

For de Finetti, the significance of his result was that "the nebulous and unsatisfactory definition of 'independent events with fixed but unknown probability,' " that is, the notion of an underlying i.i.d. probabilistic source, could be abandoned for a "simple condition of 'symmetry' in relation to our judgments of probability," that is, a property of our predictors (1937/1964,

142). In the interpretation of Braithwaite (1957) and Hintikka (1971), talk of general hypotheses, problematic from a strictly empiricist point of view, could be abandoned for constraints on methods of prediction. An allied sentiment about the dispensability of general hypotheses is expressed by Carnap (1950, 570–75) and is subsequently embraced by Solomonoff: "I liked [Carnap's confirmation function] that went directly from data to probability distribution without explicitly considering various theories or 'explanations' of the data" (1997, 76).

However, one could also reason the other way around (cf. Romeijn 2004). Namely, a representation theorem that relates a particular class of Bayesian mixtures and a particular class of predictors, like de Finetti's theorem or theorem 3, shows that this particular class of predictors operates under a particular inductive assumption. This is the inductive assumption that is codified in the priors of the Bayesian mixtures in this particular class: those patterns are assumed projectable that are represented by hypotheses that receive a nonzero prior. Thus, de Finetti's representation theorem shows that the exchangeable predictors operate under the inductive assumption of an i.i.d. source, and theorem 3 shows that the algorithmic probability predictors operate under the inductive assumption of an effective source. It is essentially this insight that defuses the argument to justify Occam's razor, as I show next.

6. Defusing the Argument. Here, I recast (sec. 6.1) and thereby defuse (sec. 6.2) the argument.

6.1. The Argument Recast. By theorem 3, the following two formulations of step 1 of the argument to justify Occam's razor are equivalent.

1. Predictors in class $\mathcal{Q}$ possess a distinctive simplicity-qua-compressibility bias.
1. Predictors in class $\mathcal{R}$ operate under inductive assumption of effectiveness.

Furthermore, since "in essentially every case" is to mean "under the assumption of effectiveness of the actual data-generating source," theorem 1 about the bound on the total risk of the algorithmic probability predictors Q_U is equivalent to theorem 2 about the consistency of the Bayesian predictors, applied to the class of effective predictors P_W^{eff}. Hence, the following two formulations of step 2 are equivalent.

2. Predictors in class $\mathcal{Q}$ are reliable in essentially every case.
2. Predictors in class $\mathcal{R}$ are consistent.

If we make the property of consistency in step 2 explicit, the two steps of the argument look as follows.

1. Predictors in class $\mathcal{R}$ operate under inductive assumption of effectiveness.
2. Predictors in class $\mathcal{R}$ are reliable under assumption of effectiveness.

Taken together, the two steps yield the conclusion that predictors that operate under the inductive assumption of effectiveness are reliable under the assumption of effectiveness.

6.2. The Argument Defused. In the original formulation, we define a class of predictors with a distinctive simplicity bias that we can subsequently prove to be reliable "in essentially every case." This formulation suggests that we have established a connection between two properties of a predictor that are quite distinct. We got out a general reliability, whereas we put in a specific preference for simplicity. This link between a simplicity bias and reliability provides an epistemic justification of Occam's razor, the principle that a predictor should have a simplicity bias.

The more explicit reformulation shows that the original formulation is misleading. We got out what we put in, after all. We define a class of predictors that operate under the inductive assumption of effectiveness, which we can subsequently prove to be reliable under the very same assumption of effectiveness.

Indeed, a renewed look at the simplicity bias described in section 3.1 unveils the notion of simplicity involved as a peculiar one. The issue of subjectivity (sec. 5.3) concretely means that we can make any finite sequence arbitrarily "simple" by an apt choice of universal Turing machine, which is the common objection against the idea that algorithmic information theory can provide an objective quantification of the simplicity of finite sequences (cf. Kelly 2008, 324–25). Then this simplicity notion could only meaningfully apply to infinite data streams, with an interpretation of "asymptotic compressibility by some machine," or, equivalently, "asymptotic goodness-of-fit of some effective hypothesis." But this notion as a property of a predictor is really the expression of a particular inductive assumption (sec. 4.2), the inductive assumption of effectiveness. The upshot is that this property is certainly a simplicity notion in the weak sense in which any inductive assumption can be seen as a specific simplicity stipulation (if only for the plain reason that an assumption restricts possibilities), but it would require a whole new argument to make plausible that the particular assumption of effectiveness is somehow preferred in defining simplicity in this sense or even gives a simplicity notion in a stronger sense. And even if it could be argued that effectiveness yields such a privileged simplicity notion, it is still not effectiveness (hence

not simplicity as such) that drives the connection to reliability: theorem 2 tells us that, at least for countable hypothesis classes, consistency holds for every inductive assumption that we formalize in W.

The conclusion is that the argument fails to justify Occam's razor.

7. Concluding Remarks. The central element of Solomonoff's theory of prediction is the constraint or assumption of effectiveness. This is clearly revealed by theorem 3, which states that Solomonoff's algorithmic probability predictors are precisely the Bayesian predictors operating under the inductive assumption of effectiveness.

The argument to justify Occam's razor does not work because the supposed connection between a predictor's simplicity preference and a predictor's general reliability, as forged by theorem 1, is really the connection between a predictor's operating under a particular inductive assumption (effectiveness, in this case) and a predictor's reliability under this same assumption. This is an instance of Bayesian consistency that is quite irrespective of the particular assumption of effectiveness.

If there exists a way to salvage the argument at all, then it would have to consist in demonstrating anew that effectiveness as an inductive assumption does lead to a fundamental simplicity notion. Regardless of the feasibility of such an undertaking, it would tie in with a more general project that certainly looks significant. This project is the inquiry into the philosophical interest of the assumption of effectiveness, particularly in the setting of sequential prediction—which, I submit, makes for the philosophical interest of Solomonoff's theory.

Now effectiveness does not appear very interesting in the naive shape of a constraint on possible data-generating sources, that is, as an assumption about processes in the world. There seems little ground for promoting the notion of effectiveness, an eminently epistemological notion that is to answer the epistemological question of what we can possibly calculate, to a constraint on the world (a positively metaphysical constraint). Nor have decades of debate about "computability in nature" uncovered support for such a move.[15]

However, the assumption of effectiveness does look very interesting in a different shape. Namely, effectiveness seems much more natural as a restriction on our own epistemic capabilities. In particular, it seems natural to say that all methods of prediction we can possibly design must be effective (sec. 2.2). If we accept this, then it is possible to prove that the algorithmic probability predictor will come to predict as well as any other pre-

15. See Piccinini (2011) for a recent overview of the debate about "physical" variants of the Church-Turing Thesis that make assertions about constraints on possible physical processes in terms of computability.

dictor. That is, an algorithmic probability predictor would represent the best we can do. This would render Solomonoff's predictor an idealized limit case of predicting at least as good as any member of a specific class of competing predictors (namely, the limit case of the class of all predictors), the central idea in the machine learning branch of universal prediction and the philosophical proposal of meta-induction. Indeed, rather than in the tradition of Carnap, addressing Hume's problem of the justification of induction by insisting on an objective starting point, this view of Solomonoff's theory is closer to a pragmatic approach to induction, going back to Reichenbach (1935).

REFERENCES

Barron, Andrew R. 1998. "Information-Theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems." In *Proceedings of the Sixth Valencia International Meeting*, ed. José M. Bernardo, James O. Berger, A. Philip Dawid, and Adrian F.M. Smith, 27–52. Oxford: Oxford University Press.

Bernardo, José M., and Adrian F. M. Smith. 1994. *Bayesian Theory*. Chichester: Wiley.

Blackwell, David, and Lester Dubins. 1962. "Merging of Opinion with Increasing Information." *Annals of Mathematical Statistics* 33:882–86.

Braithwaite, Richard B. 1957. "On Unknown Probabilities." In *Observation and Interpretation: Proceedings of the Ninth Symposium of the Colston Research Society*, ed. S. Körner, 3–11. London: Butterworths.

Carnap, Rudolf. 1945. "On Inductive Logic." *Philosophy of Science* 12:72–97.

———. 1950. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

———. 1952. *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.

Cesa-Bianchi, Nicolò, and Gabor Lugosi. 2006. *Prediction, Learning and Games*. Cambridge: Cambridge University Press.

Chaitin, Gregory J. 1969. "On the Length of Programs for Computing Finite Binary Sequences: Statistical Considerations." *Journal of the Association for Computing Machinery* 16:145–59.

Dawid, A. Philip. 1984. "Present Position and Potential Developments: Some Personal Views." *Journal of the Royal Statistical Society* A 147:278–92.

de Finetti, Bruno. 1937/1964. "La prévision: Ses lois logiques, ses sources subjectives." *Annales de l'Institut Henri Poincaré* 7:1–68. Trans. Henry E. Kyburg Jr. in *Studies in Subjective Probability*, ed. Henry E. Kyburg Jr. and Howard E. Smokler, 93–158. New York: Wiley.

Downey, Rodney G., and Denis R. Hirschfeldt. 2010. *Algorithmic Randomness and Complexity*. New York: Springer.

Gaifman, Haim, and Marc Snir. 1982. "Probabilities over Rich Languages, Testing and Randomness." *Journal of Symbolic Logic* 47 (3): 495–548.

Goodman, Nelson. 1955. *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Grünwald, Peter D. 2007. *The Minimum Description Length Principle*. Cambridge, MA: MIT Press.

Hintikka, Jaakko. 1971. "Unknown Probabilities, Bayesianism, and de Finetti's Representation Theorem." In *Proceedings of the 1970 Biennial Meeting of the Philosophy of Science Association*, ed. Roger C. Buck and Robert S. Cohen, 325–41. Dordrecht: Reidel.

Howson, Colin. 2000. *Hume's Problem: Induction and the Justification of Belief*. New York: Oxford University Press.

Hutter, Marcus. 2003. "Convergence and Loss Bounds for Bayesian Sequence Prediction." *IEEE Transactions on Information Theory* 49 (8): 2061–66.

———. 2007. "On Universal Prediction and Bayesian Confirmation." *Theoretical Computer Science* 384 (1): 33–48.

Jeffrey, Richard C. 1973. "Carnap's Inductive Logic." *Synthese* 25:299–306.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90 (420): 773–95.
Kelly, Kevin T. 1996. *The Logic of Reliable Inquiry*. New York: Oxford University Press.
———. 2008. "Ockham's Razor, Truth, and Information." In *Handbook of the Philosophy of Information*, ed. Johan F. A .K. van Benthem and Pieter Adriaans, 321–60. Dordrecht: Elsevier.
Kolmogorov, Andrey N. 1965. "Three Approaches to the Quantitative Definition of Information." *Problems of Information Transmission* 1 (1): 1–7.
Li, Ming, and Paul M. B. Vitányi. 2008. *An Introduction to Kolmogorov Complexity and Its Applications*. 3rd ed. New York: Springer.
Merhav, Neri, and Meir Feder. 1998. "Universal Prediction." *IEEE Transactions on Information Theory* 44 (8): 2124–47.
Müller, Markus. 2010. "Stationary Algorithmic Probability." *Theoretical Computer Science* 411 (1): 113–30.
Nies, André. 2009. *Computability and Randomness*. Oxford: Oxford University Press.
Ortner, Ronald, and Hannes Leitgeb. 2011. "Mechanizing Induction." In *Inductive Logic*, vol. 10 of *Handbook of the History of Logic*, ed. Dov M. Gabbay, Stephan Hartmann, and John Woods, 719–72. North-Holland: Elsevier.
Piccinini, Gualtiero. 2011. "The Physical Church-Turing Thesis: Modest or Bold?" *British Journal for the Philosophy of Science* 62:733–69.
Poland, Jan, and Marcus Hutter. 2005. "Asymptotics of Discrete MDL for Online Prediction." *IEEE Transactions on Information Theory* 51 (11): 3780–95.
Reichenbach, Hans. 1935. *Wahrscheinlichkeitslehre*. Leiden: Sijthoff.
Rissanen, Jorma J. 1989. *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.
Romeijn, Jan-Willem. 2004. "Hypotheses and Inductive Predictions." *Synthese* 141 (3): 333–64.
Schurz, Gerhard. 2008. "The Meta-inductivist's Winning Strategy in the Prediction Game: A New Approach to Hume's Problem." *Philosophy of Science* 75:278–305.
Shiryaev, Albert N. 1989. "Kolmogorov: Life and Creative Activities." *Annals of Probability* 17 (3): 866–944.
Solomonoff, Raymond J. 1960. "A Preliminary Report on a General Theory of Inductive Inference." Technical report, Zator, Cambridge, MA.
———. 1964. "A Formal Theory of Inductive Inference." Pts. 1 and 2. *Information and Control* 7:1–22, 224–54.
———. 1978. "Complexity-Based Induction Systems: Comparisons and Convergence Theorems." *IEEE Transactions on Information Theory* 24 (4): 422–32.
———. 1986. "The Application of Algorithmic Probability to Problems in Artificial Intelligence." In *Uncertainty in Artificial Intelligence*, ed. Laveen N. Kanal and John F. Lemmer, 473–91. Dordrecht: Elsevier.
———. 1997. "The Discovery of Algorithmic Probability." *Journal of Computer and System Sciences* 55 (1): 73–88.
———. 2009. "Algorithmic Probability: Theory and Applications." In *Information Theory and Statistical Learning*, ed. Frank Emmert-Streib and Matthias Dehmer, 1–23. New York: Springer.
Stalker, Douglas, ed. 1994. *Grue! The New Riddle of Induction*. Chicago: Open Court.
Suppes, Patrick. 2002. *Representation and Invariance of Scientific Structures*. Stanford, CA: CSLI.
Vitányi, Paul M. B. 2005. "Algorithmic Statistics and Kolmogorov's Structure Functions." In *Advances in Minimum Description Length*, ed. Peter D. Grünwald, In Jae Myung, and Mark A. Pitt, 151–74. Cambridge, MA: MIT Press.
Wallace, Christopher S. 2005. *Statistical and Inductive Inference by Minimum Message Length*. New York: Springer.
Wood, Ian, Peter Sunehag, and Marcus Hutter. 2013. "(Non-)equivalence of Universal Priors." In *Papers from the Ray Solomonoff 85th Memorial Conference*, ed. David L. Dowe, 417–25. New York: Springer.
Zabell, Sandy L. 2011. "Carnap and the Logic of Inductive Inference." In *Inductive Logic*, vol. 10 of *Handbook of the History of Logic*, ed. Dov M. Gabbay, Stephan Hartmann, and John Woods, 265–309. North-Holland: Elsevier.
Zvonkin, Alexander K., and Leonid A. Levin. 1970. "The Complexity of Finite Objects and the Development of the Concepts of Information and Randomness by Means of the Theory of Algorithms." *Russian Mathematical Surveys* 26 (6): 83–124.